DOCUMENT RESUME

ED 227 179 TM 830 212

AUTHOR Smith, Richard M.
TITLE A Model for Evaluating the Assessment of Partial Knowledge.
PUB DATE Mar 82
NOTE 18p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New York, NY, March 20-22, 1982).
PUB TYPE Reports - Descriptive (141) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Difficulty Level; *Evaluation Methods; Goodness of Fit; Guessing (Tests); *Models; Multiple Choice Tests; *Psychometrics; Responses; *Scoring Formulas; *Test Construction; *Test Items
IDENTIFIERS *Partial Knowledge (Tests)

ABSTRACT

There have been many attempts to formulate a procedure for extracting information from incorrect responses to multiple choice items, i.e., the assessment of partial knowledge. The results of these attempts can be described as inconsistent at best. It is hypothesized that these inconsistencies arise from three methodological problems: the difficulty of the test, the use of omits in scoring, and the fit of the items to the scoring model. This paper reviews the impact of these three problems on the assessment of partial knowledge and proposes a model which will lessen their impact on the evaluation of any partial knowledge scoring model. (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED227179

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

✓ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

A MODEL FOR EVALUATING THE ASSESSMENT OF PARTIAL KNOWLEDGE

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

R. M. Smith

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

RICHARD M. SMITH

Tm 820 212

March 1982

Paper presented at the annual meeting of the
National Council on Measurement in Education
New York, March 1982.

ABSTRACT

There have been many attempts to formulate a procedure for extracting information from incorrect responses to multiple choice items, i.e., the assessment of partial knowledge. The results of these attempts can be described as inconsistent at best. It is hypothesized that these inconsistencies arise from three methodological problems; the difficulty of the test, the use of omits in scoring, and the fit of the items to the scoring model. This paper reviews the impact of these three problems on the assessment of partial knowledge and proposes a model which will lessen their impact on the evaluation of any partial knowledge scoring model.

A Model for Evaluating the Assessment of Partial Knowledge

Throughout the last fifty years there has been a consistent belief expressed in the psychometric literature regarding the use of multiple choice items. This belief holds that the dichotomous scoring of multiple choice items does not capture the full amount of information available concerning the person's ability. Most authors have felt that the degree of incorrectness of the incorrect option chosen by the examinee is a source of additional information about the person's ability. A variety of techniques have been developed to overcome this perceived deficiency of dichotomously scoring yet retain the beneficial aspects of multiple choice items, e.g., ease of scoring, etc. For each new technique developed there have been numerous studies that have attempted to demonstrate the potential of the technique to assess partial knowledge. These demonstrations are generally based on two criteria: reliability, usually split-half or coefficient alpha; and validity, usually concurrent, estimated by correlating the test scores with some external criterion.

The literature is replete with examples of studies that show no increase in reliability or validity for the different techniques of assessing partial knowledge when compared with number right scoring or number right corrected for guessing. In those studies where increases in reliability and/or validity were reported for partial knowledge scoring, the latter required more time per item to administer. If the

traditionally-scored test had been lengthened so as to require the same amount of total time as the test using partial knowledge scoring, the increased reliability due to the increase in the number of items would offset the increase in reliability due to the use of the partial knowledge scoring. The results of these studies can best be described as inconsistent, and in fact have led many authors to suggest that the search for a method of assessing partial knowledge is futile.

It will be demonstrated that these inconsistent results are the by-products of three methodological problems which are overlooked in a majority of the studies reviewed. These problems relate to the difficulty levels of the tests, the use of omits in the data, and the problem of the fit of the items to the scoring model.

Difficulty of the Test

The first of these problems concerns the difficulty of the tests used in the studies. The successful demonstration of a procedure for assessing partial knowledge requires the presence of incorrect responses in the response matrix. The greater the number of incorrect responses, the greater the likelihood that one will be able to detect the increases in reliability and validity that are expected to accrue from the use of a technique for assessing partial knowledge.

A method of assessing partial knowledge may work, but if

one attempts to demonstrate this fact on a test where the average percent correct is 75%, it means that the technique will apply on the average to only 1 of every 4 responses. The replications necessary to detect the increase in reliability due to the application of the technique will thus be lacking.

Fewer than half of the studies involving the assessment of partial knowledge (Poizner, Nicewander, and Gettys, 1978; Hambleton, Roberts, and Traub, 1970; Patnaik and Traub, 1973; Hendrickson, 1971; Reilly and Jackson, 1973; Echternacht, 1976; Bejar and Weiss, 1977) report the average difficulty of the tests used in the studies. For those that do report the difficulties the range is from about 50% correct to 80% correct. Therefore, in the best possible case, 50% correct, the partial knowledge technique has a chance to work on only 1/2 of the items. Since most of the tests used in these studies contain 30 or less items, the best that can happen is to produce an increase in the reliability of this 15 item subset.

It also follows that the same magnitude of increase in the reliability, all other things held constant, should not be expected when two tests have different degrees of difficulty. Why should one expect the increase in reliability due to the assessment of partial knowledge for a test with an average percent correct of 75% to be the same as for a test with an average percent correct of 50% when the technique applies to 1/4 of the items in the former test and 1/2 of the items in the

latter? Despite this, not one of the authors reviewed mentioned the dependency of the maximum increase in reliability on the average difficulty of the test, nor its possible effect on significance testing. In fact, only half of the authors considered the average item difficulty important enough to report in the article, thus making comparison across studies on this point difficult.

Most of the tests used in these studies, however, were pre-existing tests, usually standardized, and built under the usual guidelines for dichotomously scored multiple choice tests. It is, therefore, safe to assume that the average difficulty of these tests ranged between 50% and 80% correct, quite similar to those difficulties which are reported.

Omits

The second methodological issue relates to the disposition of omitted responses in the data matrix and is most apparent in studies using the empirical response weighting procedure (Hendrickson, 1971; Reilly and Jackson, 1973). This procedure uses either the frequency of responses to various options or the overall abilities of the persons selecting the various options to calculate their weights. The use of this procedure has led to several incongruities. In some instances, higher scores were assigned to incorrect options than were assigned to the correct option. In others, higher scores were assigned to individuals who left an item blank than to those who chose one

of the incorrect options.

A typical response of authors to these anomalies is given by Cross and Frary (1978):

> In only one sample and for one question was the weight of the correct choice lower than the weight for a distractor and in this instance the difference was small(.06). These results suggest that the choice weights obtained using groups of 211 examinees were reasonably stable and with double crossed validation, reasonable confidence can be placed in the results of choice weighting procedures described herein.

How it can be reasonable for an incorrect option to carry a higher weight than the correct answer to the item is puzzling, but this lack of concern typifies the work of the authors using the empirical response weighting technique. Since these authors had no trouble weighting incorrect responses higher than correct responses, it is easy to see why they had little conceptual trouble with omitted responses carrying higher weights than incorrect responses.

By the mid-1970's, researchers began to realize that the inclusion of omitted responses in empirical response weighting studies tended to increase the reliability of the test scores but decreased the validity of those scores. It was hypothesized that the empirical weighting technique capitalized on a reliable but unrelated behavior, i.e., the tendency to omit responses. This hypothesis brought about a new set of studies that tried to eliminate the effects of omitting. Reilly (1975) calculated empirical weights and then assigned a weight equal to the average weight of all options to the omitted response.

He reasoned that this score would have been the person's average score if he had guessed on all omitted items. Cross and Frary(1978) entered a random response for all omitted responses in their data and then calculated the empirical weights. Hakstain and Kansup (1975) and Kansup and Hakstain (1975) deleted all persons with omitted responses from their studies which, despite allowing extra time and instructing students to answer all items, required the deletion 10% of their subjects.

Despite the seemingly pervasive nature of omitting responses while taking tests, the issue of omits is seldom addressed in studies involving techniques other than weighted response. In the typical elimination scoring procedure the score of zero, commonly given to omitted responses, is equal to the average score a person would receive if he had eliminated one response at random. This has the same effect on the total score as Reilly's assigning the average option weight to omitted responses.

It is unclear from the probabilistic scoring literature exactly how omits were handled, or if, in fact, they exist. In some studies involving the answer-until-correct (AUC) technique a value of zero is assigned when an item is omitted. Since the expected score due to guesing would be one on a four option item, this value is less than the expected score for guessing one response per item.

In the Hanna study (1975) the AUC score was calculated by

subtracting the total number of responses made in finding the correct answers, as determined by erasures, from the total number of possible responses. A perfect score in this case would be (mL)-L , where "m" is the number of choices per item and "L" is the number of items. However, if a person had not attempted an item at all his score for that item, according to the description given in the study, would have been "m" rather than the "m-1" he would have gotten if he answered it correctly on the first attempt. This is clearly an unacceptable result, but Hanna fails to state how omits, if any, were handled in his study.

Despite the variety of methods used to correct for the presence of omits in the data, the problem has not been solved. Every procedure developed either leads to a logical contradiction or is based on an arbitrary rule as to what the person would have scored if in fact he had answered the item.

Model Fit

The third methodological issue concerns the fit of the items to the psychometric model on which the scoring procedure for that test is based. One characteristic of the studies reviewed is that they all involved the comparison of dichotomous number right scoring with some form of polychotomous scoring. To make this type of comparison most authors took a pre-existing test and administered it under two experimental conditions, that of dichotomous scoring and the

particular partial knowledge technique under study.

Although it is never stated explicitly, two assumptions seem to have been made in these studies. First, it is assumed that the tests were designed to fit the requirements of the dichotomous classical psychometric model. This implies that the available item pool was edited to remove misfitting items . Items can be judged as misfitting on the basis of two properties, difficulty and discrimination. The item pool is usually edited to remove easy and difficult items. Final test construction typically involves selecting items such that the average percent correct for the pre-tested items is .65 (for 4 choice items). The item pool is also edited to remove items with low point biserial correlations. Items with point biserials lower than .20 to .30 are usually deleted. Editing the items in this fashion increases the number of items that fit the psychometric model and thus increases the reliability of the test scores.

The second assumption implicit in using pre-existing tests for comparing the reliability and validity of techniques of assessing partial knowledge is that items that are designed to fit the dichotomous model will also fit the partial knowledge model. Although it has been a standard procedure to pretest items prior to use in dichotomous tests to insure the fit of the items to the model, and although ongoing programs of item review to check the continued fit of the items to the

dichotomcus model are common, not one of the studies reviewed checked the appropriateness of the items to the partial knowledge scoring model prior to the start of the study. Only one of the studies (Rippey 1968) checked the fit of the items to the partial knowledge model after the testing.

In his study on probabilistic scoring Rippey (1968) reports

> ... further increases in reliability (for probabilistic scoring) could have been obtained by appropriate item selection. An item analysis showed that items which had low or negative correlations with the total test score were not the same items when the test was administered in a conventional way.

The implication of this statement is clear. Fit to the dichotomous model is not a sufficient condition for fit to the polychotomous model. Since Rippey gives no indication of the number of items that fit the dichotomous model but did not fit the polychotomous model, it is difficult to gauge the extent of the problem. However, since test reliability can be thought of as a function of the item-total score correlation, it is clear that the presence of a few items with the negative correlations that Rippey reports could be enough to mask the gain produced by the items that did fit both models.

Consider the example of a twenty-item test designed for use as a dichotomously-scored instrument. Suppose that 15 of the 20 items on the test are appropriate as polychotomous items. If the test administered in its dichotomous and in

polychotomous forms have the same reliability, as is usually the case reported in the literature, our first conclusion might be that the use of this particular method of assessing partial knowledge does not increase the reliability of the test. This conclusion ignores the fact that 5 of the 20 items on the test did not work in the polychotomous case. Thus, they were not contributing to the overall increase in reliability expected for polychotomous scoring. If we deleted these 5 items, with little loss in reliability, and adjusted for the differences in test length the polychotomous test would then be more reliable. Given our knowledge of the problems of shifting items from dichotomous to polychotomous scoring, this methodological issue is a third possible explanation of the inconsistency in the results of partial knowledge scoring reported to date.

Summary

The three methodological issues discussed here are not separate when it comes to their impact on the results reported in studies assessing the measurement of partial knowledge. They tend to compound each other. This is particularly true of difficulty and fit. Even if one uses a relatively difficult test by dichotomous standards, e.g., an average of 50% correct, this means that on the average only 1/2 of the items can contribute to the assessment of partial knowledge. If under this condition only 1/2 of the items fit the polychotomous scoring model, and experience indicates that this figure can be as low as 1/3 for standardized tests, then only an average of

1/4 of the items will be contributing to increasing the reliability of the test under polychotomous scoring. If these figures are as typical as the literature suggests, the best to be expected is that only an average of 1/4 of the items will contribute to the assessment of partial knowledge. It is no wonder that increases in reliability and validity due to polychotomous scoring are rare.

Proposed Model

It would appear that there are ways of overcomming the methodological problems discussed above. The item difficulty problem is fairly easy to deal with given existing item banks. We can build a test with an expected percent correct in the 35% to 45% range. An alternative suggestion is to use a pre-existing test with an average percent correct of more than 50%, but to focus the validity and reliability calculations on those individuals who had less then 40% of the items correct. Either alternative would assure that at least the majority of the people would have more than 50% of the items incorrect so that there would be the possibility of evaluating partial knowledge demonstrated by their incorrect responses. For those concerned about guessing, the use of a strong psychometric model, such as the Rasch model, would allow for the detection of examinees who guessed due to the difficulty of the test. They can then be eliminated from the calculation of reliability and validity since their responses would not be a valid indication of their ability.

The problem of the tendency of certain people to omit items, which apparently confounds more than just empirically-derived response weights, can also be handled with the Rasch measurement model. In this psychometric model, a person's ability estimates are independent of the items to which that person respondeo. This property allows us to estimate ability solely on the basis of the items that a person attempted, thus freeing the ability estimates from their interaction with the tendency to omit. This can remove the confounding influence of this tendency from the reliability and validity calculations.

Finally, the problem of the fit of the items to a scoring model can be solved by pretesting the items in both polychotomous and dichotomous scoring models. On the basis of the results we should be able to build a test using only items that fit both models. This may be accomplished using the Cronbach-Merwin procedure (1955) or with some other test of fit, for example, ordered step values in the Rasch polychotomous model. This would not guarantee that all of the items would work in the final form, since these tests of fit are sample-dependent, but we should be able to write tests where over 90% of the items fit the polychotomous model. We can then focus the final analysis on the items that fit the model.

In combination these three steps will allow us to free the assessment of partial knowledge from the confounding factors

that have been discussed above. If we were to construct a test in this fashion we would expect that, on the average, 2/3 of the items would be contributing to the assessment of partial knowledge, rather than the 1/4 that we currently find in the literature. This should greatly increase the chances of finding significant increases in reliability and validity where they exist.

REFERENCES

Arnold, J.C. and Arnold, P.L. On scoring multiple choice exams allowing for partial knowledge. The Journal of Experimental Education, 1970, 39, 8-13.

Bejar, I.I., and Weiss, D.J. A comparison of empirical differential option weighting scoring procedures as a function of inter-item correlations. Educational and Psychological Measurement, 1977, 37, 335-340.

Claudy, J.G. Biserial weights: A new approach to test item option weighting. Applied Psychological Measurement, 1978, 2, 25-30.

Collet, L.S. Elimination scoring: An empirical evaluation. Journal of Educational Measurement, 1971, 8, 209-214.

Cronbach, L.J. and Merwin, J.C. A model for studying the validity of multiple-choice items. Educational and Psychological Measurement, 1955, 15, 337-352.

Cross, L.H. and Frary, R.B. Empirical choice weighting under "guess" and "do not guess" directions. Educational and Psychological Measurement, 1978, 38, 613-620.

Ecternacht, G. Reliability and validity of item option weighting schemes. Educational and Psychological Measurement, 1976, 36, 301-309.

Hakstain, A.R. and Kansup, W. A comparison of several methods of assessing partial knowledge in multiple-choice tests: II. testing procedures. Journal of Educational Measurement, 1975, 12, 231-239.

Hambleton, R.K., Roberts, D.M., and Traub, R.E. A comparison of the reliability and validity of two methods on assessing partial knowledge on multiple choice tests. Journal of Educational Measurement, 1970, 7, 75-82.

Hanna, G.S. Incremental reliability and validity of multiple-choice tests with an answer-until-correct procedure. Journal of Educational Measurement, 1975, 12, 175-178

Hansen, R. The influence of variables other than knowledge on probabilistic tests. Journal of Educational Measurement, 1971, 8, 9-14.

Hendrickson, G.F. The effects of differential option weighting on multiple-choice objective tests. Journal of Educational Measurement, 1971, 8, 291-296.

Kansup, W. and Hakstain A.R. A comparison of several methods of assessing partial knowledge in multiple-choice tests: I. scoring procedures. Journal of Educational Measurement, 1975, 12, 219-229.

Merwin, J.C. Rational and mathematical relationships of six scoring procedures applicable to three-choice items. Journal of Educational Psychology, 1959, 50, 153-161.

Patnaik, D., and Traub, R.E. Differential weighting by judged degree of correctness. Journal of Educational Measurement, 1973, 10, 281-286.

Poizner, S.B., Nicewander, W.A. and Gettys, C.F. Alternative response and scoring methods for multiple choice items: an empirical study of probabilistic and ordinal response modes. Applied Psychological Measurement, 1978, 2, 83-96.

Reilly, R.R. Empirical option weighting with a correction for guessing. Educational and Psychological Measurement, 1975, 35, 613-619.

Reilly, R.R., and Jackson, R., Effects of empirical option weighting on reliability and validity of an academic aptitude test. Journal of Educational Measurement, 1973, 10, 185-193.

Rippey, R. Probabilistic testing. Journal of Educational Measurement, 1968, 5, 211-215.

Rippey, R. A comparison of five different scoring functions for confidence tests. Journal of Educational Measurement, 1970, 7, 165-170.

Sabers, D.L. and White, G.W. The effect of differential weighting of individual item responses on the predictive validity of an aptitude test. Journal of Educational Measurement, 1969, 6, 93-96.

Wood. R., Multiple choice: A state of the art report. Evaluation in Education, 1977, 1, 231-239.